Università degli studi di
Milano-Bicocca

Data management and visualization

Final Project

Twitter può influenzare i mercati finanziari? Caso studio: Pfizer, AstraZeneca e Sinopharm.

Authors:
Christian Internò - 876238 - c.interno@campus.unimib.it
Salvatore Vizzi - 864010 - s.vizzi@campus.unimib.it
Martina Gulino - 864027 - m.gulino3@campus.unimib.it

Abstract

Il lavoro è consistito nel capire se i tweet pubblicati riferiti ad una particolare azienda influenzano l'andamento in borsa dell'azienda. Per rispondere a questa domanda ci siamo interessati a tre diverse aziende farmaceutiche che hanno pordotto vacini contro il COVID19: Sinopharm, AstraZeneca, Pfizer. Sono quindi stati estratti i tweet contenenti il nome di almeno una delle tre aziende integrandoli con i dati azionario. Le visualizzazioni ci hanno consentito di rappresentare gli andamenti giornalieri e orari del numero di tweet e degli andamenti in borsa, includendo anche ulteriori parametri quali la sentiment e gli engagement dei tweet. Le infografiche non ci hanno consentito di dare un giudizio chiaro per quanto riguarda un possibile nesso tra i due andamenti. Il software utilizzato per la parte di integrazione dei dati è Python Notebook. Per le visualizzazioni grafiche è stato utilizzato Tableau.

1 Introduzione

Nell'ultimo anno, la pandemia COVID-19 ha profondamente influenzando la nostra vita quotidiana imponendo un nuovo modo di concepire il lavoro, l'istruzione e portando grandi cambiamenti nel nostro modo di rapportarci con le persone che ci circondano. Dopo numerosi mesi di restrizioni e distanziamenti, la vaccinazione sta permettendo il lento ritorno ad una vita normale. Infatti, grandi studi clinici e sviluppi hanno permesso la creazione di vaccini per la lotta contro il Covid19 da parte di diverse case farmaceutiche. Negli ultimi mesi, la vaccinazione di massa ha permesso di tornare gradualmente alla normalità. Ma se i vaccini stanno portando a grandi miglioramenti all'interno delle nostre vite, le nostre azioni, i nostri discorsi, le nostre idee, i post che abitualmente pubblichiamo, influenzano le quotazioni delle grandi aziende produttrici di vaccini?

Per rispondervi sono stati estratti e analizzati i post pubblicati sulla piattaforma Twitter e il valore delle azioni delle principali case farmaceutiche che giocano un ruolo fondamentale nella lotta mondiale contro la pandemia attraverso la larga distribuzione dei loro vaccini: Sinopharm, Pfizer e AstraZeneca. Queste hanno suscitato il nostro interesse poiché provengono da regioni geografiche e mercati diversi e partecipano a dinamiche geopolitiche complesse ed interessati nel mondo globalizzato attuale: Sinopharm, impresa statale cinese le cui azioni sono quotate nella borsa di Hong Kong,

Pfizer grande leader mondiale nel settore della ricerca, della produzione e della commercializzazione di farmaci le cui azioni sono quotate nella borsa di New York ed infine AstraZeneca multinazionale biofarmaceutica anglo-svedese anch'essa quotata a New York.

La tesi che vogliamo confutare è dunque: i tweet postati relativi a ciascuna azienda influenzano l'andamento nei mercati finanziari? Il lavoro si è suddiviso in vari passi: la raccolta dati provenienti da due fonti diverse, i successivi preprocessing e integrazione dei dati, ed infine la loro analisi accompagnata da visualizzazioni infografiche.

Per il primo passo di raccolta e processing dei dati, il lavoro si è suddiviso in due parti ben distinte: la prima riguardante la raccolta dei dati tramite web scraping e query api, la seconda ha permesso la raccolta dati e la successiva integrazione in real time attraverso i tools Kafka e Nifi.

2 Data Management

I requisiti che abbiamo deciso di soddisfare per rispondere alla nostra domanda sono la varietà e la velocità. Per soddisfare il primo requisito sono stati integrati dataset provenienti da due fonti diverse, in particolare Twitter e YahooFinance. Per quanto riguarda la velocità è stata costruita un'architettura che consentisse di gestire l'acquisizione del dato in real-time attraverso la tecnologia middleware "Apache Kafka".

2.1 Data gathering

La prima parte del lavoro ha riguardato l'acquisizione e l'integrazione dei dati.

2.1.1 Raccolta dati

La prima fase del progetto ha visto la raccolta dei dati relativi ai titoli azionari attraverso la libreria denominata "yfinance", creata in seguito alla disattivazione dell'API di "Yahoo!Finance" e consentendo di raccogliere dati storici di mercato. Le query API sono state effettuate a cadenza settimanale e si è notato come i dati relativi al prezzo dei titoli possedevano una granularità molto elevata essendo valori aggiornati regolarmente al minuto.

I dati storici raccolti contenevano il prezzo di apertura e di chiusura della borsa, il prezzo più alto e più basso registrato durante la giornata, il volume del titolo azionario ovvero il numero delle transazioni effettuate durante un certo periodo di tempo per un determinato titolo. Quando un contratto viene scambiato vi sono sempre due controparti, un compratore e un venditore, che effettuano una transazione. Ogni transazione va ad incrementare il volume dell'operazione di trading. Bisogna però tenere presente che il volume viene calcolato in base al numero dei contratti scambiati e l'ammontare delle transazioni effettuate è da escludere.

Attraverso questa libreria sono stati raccolti dati per le aziende Sinopharm, Pfizer e Astrazeneca durante il periodo dall'8 Aprile 2021 sino al 21 Maggio 2021.

Per i dati relativi ai tweet è stata utilizzata una libreria denominata "snscrape", scraper per i social network che permette l'estrazione di dati per varie piattaforme social (ad es. Facebook, Instagram, Twitter ecc.). Ai fini del nostro lavoro, è stata utilizzata per la raccolta di posto contenenti le parole chiavi relative alle tre diverse aziende e con il vincolo che fossero scritti in lingua inglese. Il periodo selezionato risulta il medesimo di quello considerato per i dati finanziari (dall'8 Aprile fino al 21 Maggio).

Per ogni post si sono ottenuti i campi relativi alla data e all'ora a cui è stato pubblicato il tweet, al numero totale di like, al conteggio dei retweet, al numero di commenti, ad informazioni riguardanti l'autore del post quali il suo numero di followers ed infine in certi casi anche informazioni riguardante la localizzazione in cui è stato pubblicato il tweet.

Una prima osservazione che si può fare è che, seppur relativi allo stesso arco temporale, la numerosità dei post relativi a ciascuna azienda è diversa. Per Sinopharm si sono raccolti circa 50 mila post. Per Pfizer i post pubblicati hanno superato il milione e per l'azienda anglo-svedese si sono raccolti oltre mezzo milione di tweets.

Per ogni tweet è stata effettuata una Sentiment Analysis, tecnica che rileva il sentimento sottostante in una parte di testo, classificandosi come positivo, negativo o neutro. Si è utilizzata a la libreria "vaderSentiment" che permette di associare un valore numerico relativo alla positività, neutralità e negatività del testo analizzato.

2.1.2 Preprocessing e integrazione

Una prima considerazione da fare è che gli orari dei tweet sono riferiti all'orario del meridiano di Greenwich. Invece gli orari dei dati azionari recuperati dal sito Yahoo Finance corrispondono agli orari di apertura del varie borse: orario della borsa di New York per i dati di Pfizer e Astrazeneca e orario della borsa di Hong Kong per Sinopharm. Sono state prima di tutto effettuate delle trasformazioni per riportare tutti i dati al fuso orario di Greenwich. Il preprocessing dei dati e la loro integrazione è stata divisa in due diverse fasi.

Poiché i dataset ottenuti da ciascuna fonte di dati possiedono una frequenza temporale diversa si è deciso di procedere con due tipi di integrazioni. Infatti, i tweet risultano continui nel tempo mentre i dati relativi al valore delle azioni vengono registrati ogni ora durante gli orari di apertura della borsa: dal lunedì al venerdì dalle 9:30 alle 16:00 per ciascuna delle due borse considerate.

Considerata la domanda a cui si vuole dare risposta, e noto il vincolo dell'orario di apertura e chiusura della borsa, si è deciso di effettuare una doppia analisi distinguendo l'integrazione in due fasce: quella oraria e quella giornaliera. Di conseguenza, l'integrazione oraria viene applicata durante l'orario di apertura della borsa, andando a selezionare i tweet postati all'ora t e associandolo al corrispondente valore delle azioni all'ora t+1.

Figure 1: Disegno dell'integrazione oraria

Figure 2: Disegno dell'integrazione giornaliera

Quindi prendendo in considerazione il fatto che la borsa è aperta dalle 9:30 alle 16, i tweet sono stati raggruppati per ora dalle ore 8:30 alle ore 15 e rispettivamente associati agli andamenti in borsa delle ore 9:30 alle ore 16.

Quella giornaliera invece, viene applicata durante gli orari di chiusura della borsa. Per quest'ultima, vengono selezionati i tweet che sono stati scritti dall'orario di chiusura del giorno t fino all'orario di chiusura del giorno t+1 associandoli al prezzo dell'orario di chiusura del giorno t+1. Per quanto riguarda la chiusura della borsa nei giorni del sabato e della domenica, si è deciso di pesare i tweet per ogni fine settimana ai fini dell'analisi per evitare un accumulo sproporzionato dei tweet scritti nel weekend.

Una volta integrati i dati, per entrambe le tipologie di integrazione oraria e giornaliera, sono stati calcolati vari fattori, tra cui il numero di tweet postati, il massimo numero di like, la media pesata del numero di like, il conteggio dei retweet e dei reply ed infine la media della sentiment dei tweet.

2.2 Real time

Per affrontare il vincolo della velocità, la raccolta dati e successiva integrazione è stata riformulata.

2.2.1 Raccolta dati

Per la raccolta dati effettuata in real time infatti sono state utilizzate diverse tecnologie. Per l'acquisizione dei dati in tempo reale relativi alle azioni dal sito "Yahoo! Finance" attraverso lo scraping, è stata utilizzata una tecnologia

asincrona come Apache Kafka che permette di disaccoppiare l'acquisizione dalla successiva memorizzazione del dato. Sono stati utilizzate tre topics diversi relativi alle tre aziende diverse.

Pertanto, lo scraping raccoglie gli stessi dati acquisiti precedentemente ma in tempo reale ed essi vengono successivamente memorizzati su MongoDB.

Per automatizzare il processo di scraping è stato utilizzato uno schedulatore che permettesse di raccogliere i dati a cadenza oraria impostando gli orari di apertura della borsa (in base ovviamente ai vari fusi orari).

Per la raccolta dei tweet, invece, è stata utilizzata la libreria Tweepy raccogliendo i dati e memorizzando su MongoDB. Anche qui è stato utilizzato Kafka per disaccoppiare l'acquisizione dalla memorizzazione del dato. Per monitorare il flusso di tweet è stato utilizzata la tecnologia di Nifi creando anche in questo caso tre topics diversi relativi alle tre diverse aziende.

Questo sistema di monitoraggio permette di conteggiare il numero di tweet che vengono raccolti e di valutare il corretto funzionamento dell'infrastruttura analizzando anche il numero di tweet che vengono perduti.

2.2.2 Preprocessing e integrazione

Infine, anche l'integrazione è stata automatizzata, i due tipi di integrazione precedentemente spiegati sono stati adattati per recuperare i dati precedentemente memorizzati sul database MongoDB, effettuare le operazioni di preprocessing e di integrazione, e successivamente riscriverli su database. Anche in questa parte, è stato utilizzato uno schedulatore che permettesse di effettuare l'integrazione oraria nel caso in cui il codice venisse eseguito durante l'orario di apertura della borsa e l'integrazione giornaliera durante gli orari di chiusura della giornata borsistica.

3 Data Visualization

Una volta effettuate le fasi di raccolta, preprocessing ed integrazione dei dati, si è proceduto con l'analisi dei dati e la creazione di visualizzazioni grafiche attraverso l'utilizzo del software Tableau. Sono state riportate solamente le visualizzazioni riguardante l'azienda Pfizer.

3.1 Prima infografica: numero di tweet effettuati giornalmente e variazioni del prezzo azionario

Nella prima visualizzazione si è deciso di mostrare tramite un grafico ad albero e scala di colore il numero di tweet effettuati relativi ad una data azienda. Il colore varia da un blu scuro ad un verde acqua che indica rispettivamente la giornata in cui sono stati pubblicati il maggior numero di tweet e quella in cui sono stati pubblicati il minor numero di tweet.

I quadrati sono disposti in ordine decrescente rispetto alla diagonale che parte dall'alto a sinistra per arrivare in basso a destra. Si evince che la giornata con più tweet pubblicati rispetto a Pfizer sia il 15 aprile 2021. In corrispondenza di questa data, Pfizer annunciava che era necessaria la terza dose per la lotta contro il virus.

Per avere una visione più completa, si è deciso di unire l'infografica ad albero appena descritta con il grafico nel quale le linee ordinate in modo crescente rappresentano la variazione percentuale del prezzo delle azioni tra due giorni successivi. L'obiettivo è di associare a ciascun giorno il numero di tweet pubblicati e la variazione percentuale del prezzo riguardante il titolo azionario.

Osservando quindi parallelamente i due grafici si può dedurre come nel giorno in cui si è registrato il massimo numero di tweet si associa una variazione percentuale del prezzo negativa nel successivo rispetto a quello precedente del 2.5%.

Per quanto riguarda AstraZeneca e Sinopharm, i giorni in cui si vede il maggior numero di tweet sono il 9 aprile e il 10 maggio rispettivamente. Il 9 aprile, l'agenzia europea delle medicine affermava che c'era un rischio molto basso di trombosi per il vaccino di AstraZeneca. In corrispondenza di questa data, si osserva una variazione del prezzo del 0.57%. Per Sinopharm, si registra una variazione del prezzo pari a -2.5% in corrispondenza del 10 maggio.

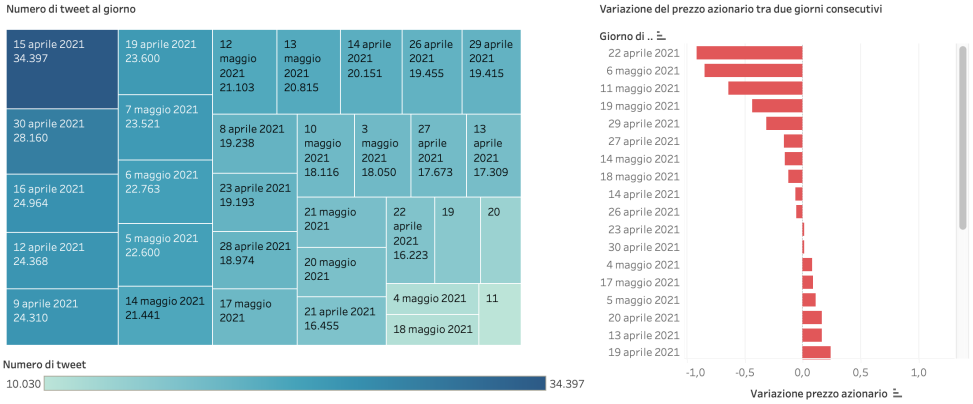


Figure 3: In corrispondenza di che giorno sono associati un dato numero di tweet una data variazione del prezzo delle azioni?

3.2 Seconda infografica: numero di tweet e valore del titolo azionario con integrazione oraria

Per la seconda visualizzazione si è deciso di mostrare gli andamenti orari (riferiti quindi agli orari di apertura della borsa) del numero totale di tweet scritti in riferimento alle singole aziende (evidenziato in blu) e della rispettiva variazione del titolo azionario (evidenziata in rosso).

La visualizzazione interattiva permette di scegliere quale azienda osservare. È possibile selezionare la settimana di interesse che si vuole considerare. Le ore corrispondono agli orari di apertura della borsa di riferimento rispetto al meridiano di Greenwich. L'infografica mostra diversi parametri per ciascun dato: il numero di tweet, del prezzo del titolo, l'engagement dei tweet (somma di like, retweet e reply dei tweet) ed infine i riferimenti temporali del dato che stiamo considerando.

Per la visualizzazione oraria relativa a Pfizer, si osserva come il 28 aprile alle ore 15:00 vi sia una corrispondenza positiva tra la variazione del prezzo del titolo azionario e il numero di tweet effettuati.

Per Sinopharm, si osservano picchi positivi del numero di tweet e della variazione del valore azionario in corrispondenza della stessa ora il 22 aprile.

Anche per quanto riguarda AstraZeneca, non apparre nessuna chiara corrispondenza tra i due andamenti.

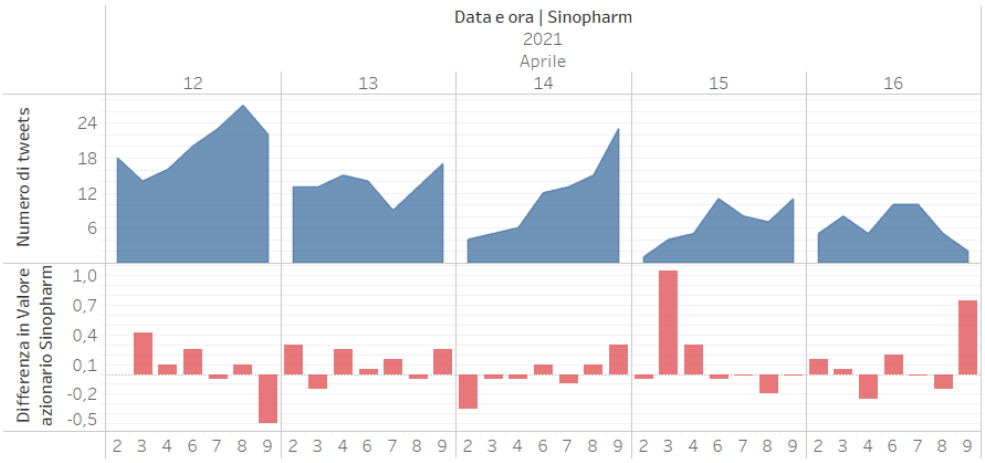


Figure 4: Si possono confrontare gli andamenti delle variazioni orarie del numero di tweet e la variazione del valore delle azioni?

3.3 Sentiment analysis dei tweet e andamento del titolo in borsa con integrazione giornaliera

Per la terza infografica, si è voluto aggiungere all'andamento giornaliero del numero di tweet anche le caratteristiche di sentiment e di numero di like dei tweet pubblicati in un determinato giorno. La sentiment è stata rappresentata tramite il colore della curva superiore dato in base alla positività (blu) e negatività (rosso) del messaggio in referimento all'azienda. Il diverso spessore della curva indica invece il diverso numero di like. Nella parte inferiore della visualizzazione vi è la curva relativa alla variazione del valore del titolo in borsa in riferimento al prezzo di chiusura di ogni giornata.

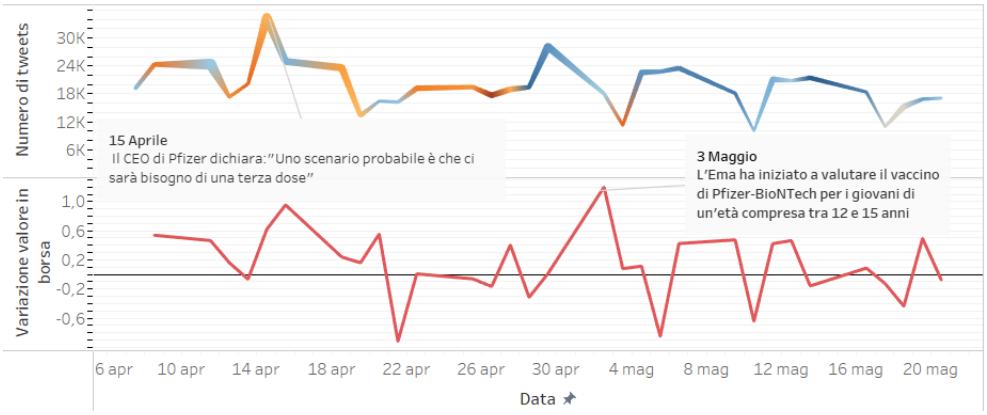


Figure 5: Si possono confrontare gli andamenti dei tweet e della variazione del prezzo delle azioni riferiti ad una data azienda?

Per quanto riguarda l'azienda Pfizer si può osservare come tra il 30 aprile e il 3 maggio si registra una variazione positiva del prezzo riferito al titolo in questione, in corrispondenza di una forte variazione del numero di tweet e di una sentiment positiva rappresentata con colore blu. In tale periodo l'EMA ha iniziato a valutare l'uso del vaccino Pfizer per giovani in un'età compresa tra i 12 e 15 anni.

Per AstraZeneca si osserva un picco per il numero di tweet tra il 6 e 9 aprile in corrispondenza di un importante numero di like, di una sentiment negativa e di un andamento negativo del valore delle azioni. In questo periodo il comitato per la sicurezza dell' EMA aveva concluso che i coaguli di sangue dovevano essere elencati come effetti molto rari per il vaccino Astrazeneca.

Anche per quanto riguarda Sinopharm si osserva tra il 9 ed il 12 maggio un aumento considerevole del numero di tweet in corrispondenza di un aumento generale di like, una sentiment analysis negativa e quindi un andamento negativo del valore del titolo. In quei giorni l'OMS approva il vaccino Sinopharm per l'uso di emergenza.

3.4 Correlazione tra il numero di tweet pubblicati giornalmente e valore del prezzo del titolo azionario

Infine, tramite l'ultima infografica si è voluto rappresentare se vi è correlazione tra il numero di tweet pubblicati giornalmente e il valore del prezzo

del titolo azionario. Si questa analisi si propone un grafico a dispersione (Scatter plot).

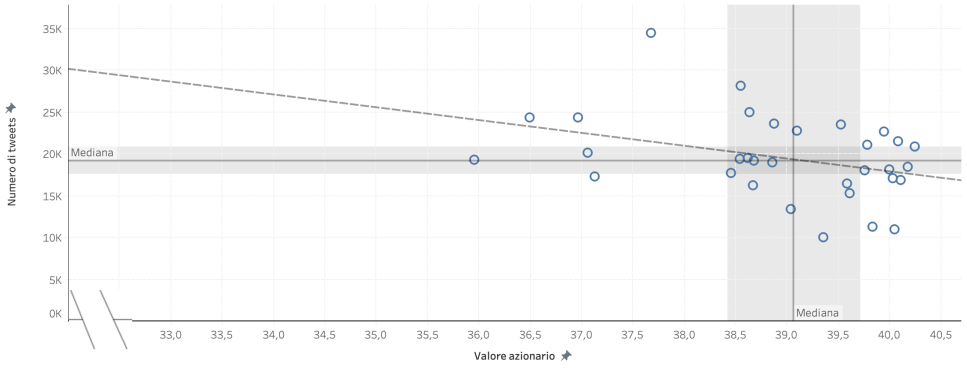


Figure 6: Esiste una correlazione tra il numero di tweet pubblicati riferiti ad una data azienda e il valore del prezzo dell'azione della stessa azienda?

La visualizzazione mostra sull'asse orizzontale il valore del prezzo dell'azione di un'azienda e sull'asse verticale il numero di tweet pubblicati riferiti alla stessa azienda. La linea tratteggiata grigia è una linea di tendenza lineare che permette di capire la tendenza dei punti rappresentati. Oltre a questa sono state rappresentate la mediana del numero di tweet (linea orizzontale) e la mediana del prezzo dell'azione (linea verticale) con i rispettivi range interquartili (aree grigie) per individuare dove sono posizionati i punti rispetto ad esse. Questa visualizzazione permette anche di evidenziare se ci sono outlier ovvero punti anomali distanti dalle altre osservazioni disponibili. È possibile scegliere se si vogliono visualizzare dati orari o dati giornalieri. Considerando l'azienda farmaceutica Pfizer si può osservare come all'aumentare del numero totale di tweet pubblicati si nota una variazione di prezzo negativa.

Anche per AstraZeneca si osserva una correlazione negativa, poiché all'aumentare il numero di tweet totali effettuati diminuisce il valore del titolo azionario.

Al contrario, per quanto riguarda l'azienda cinese Sinopharm vi è correlazione positiva, con l'aumentare del numero di tweet il valore del titolo aumenta.

4 Valutazione Infografiche

4.1 User Test

Per la parte di Assessement sono state individuate 6 domande (due per ogni infografica) da sottoporre a 6 persone. Le domande in questione sono state:

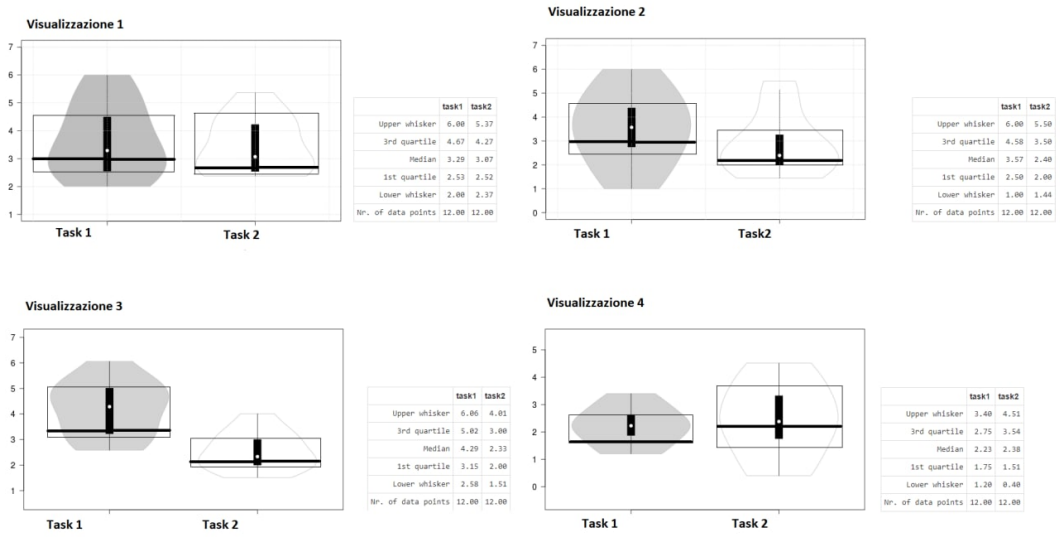
- Prima infografica
    1. In corrispondenza di che data si ha la sentiment più alta?
    2. In corrispondenza di che giorno si è avuto il minimo numero di like?
- Seconda infografica
    1. In corrispondenza di che giorno si è avuto il picco relativo al valore del titolo azionario?
    2. In corrispondenza di che giorno si è avuto il picco relativo al numero di tweet effettuati?
- Terza infografica
    1. In che giorno si è avuta la massima variazione del prezzo corrispondente al massimo numero di Tweet?
    2. Nel giorno in cui si è avuto il massimo numero di Tweet che variazione di prezzo si è avuta?
- Quarta infografica
    1. Aumentando il numero dei tweet effettuati tendenzialmente il valore del prezzo di chiusura aumenta o diminuisce?
    2. Quanti outlier (punti anomali) sono presenti?

Di seguito sono riportati i violin plot relativi ai tempi di risposta per ciascuna domanda e gli stacked bar chart relativi agli errori compiuti da un campione di 6 intervistati. I tempi di risposta ottimali sono stati individuati facendo compilare ad ogni lettore il questionario 3 volte in modo veloce e lento, calcolando quindi la media.

Si presentano le principali difficoltà nel rispondere alle domande:

1. Vi sono state difficoltà relative alla comprensione e nell'utilizzo dei filtri.

2. Difficoltà nel comprendere il numero preciso di outliers presenti nel grafico a dispersione.



	task1	task2
Upper whisker	6.00	5.37
3rd quartile	4.67	4.27
Median	3.29	3.07
1st quartile	2.53	2.52
Lower whisker	2.00	2.37
Nr. of data points	12.00	12.00

	task1	task2
Upper whisker	6.00	5.50
3rd quartile	4.58	3.50
Median	3.57	2.40
1st quartile	2.50	2.00
Lower whisker	1.00	1.44
Nr. of data points	12.00	12.00

	task1	task2
Upper whisker	6.06	4.01
3rd quartile	5.02	3.00
Median	4.29	2.33
1st quartile	3.15	2.00
Lower whisker	2.58	1.51
Nr. of data points	12.00	12.00

	task1	task2
Upper whisker	3.40	4.51
3rd quartile	2.75	3.54
Median	2.23	2.38
1st quartile	1.75	1.51
Lower whisker	1.20	0.40
Nr. of data points	12.00	12.00

Figure 7: Violin plot del tempo di risposta

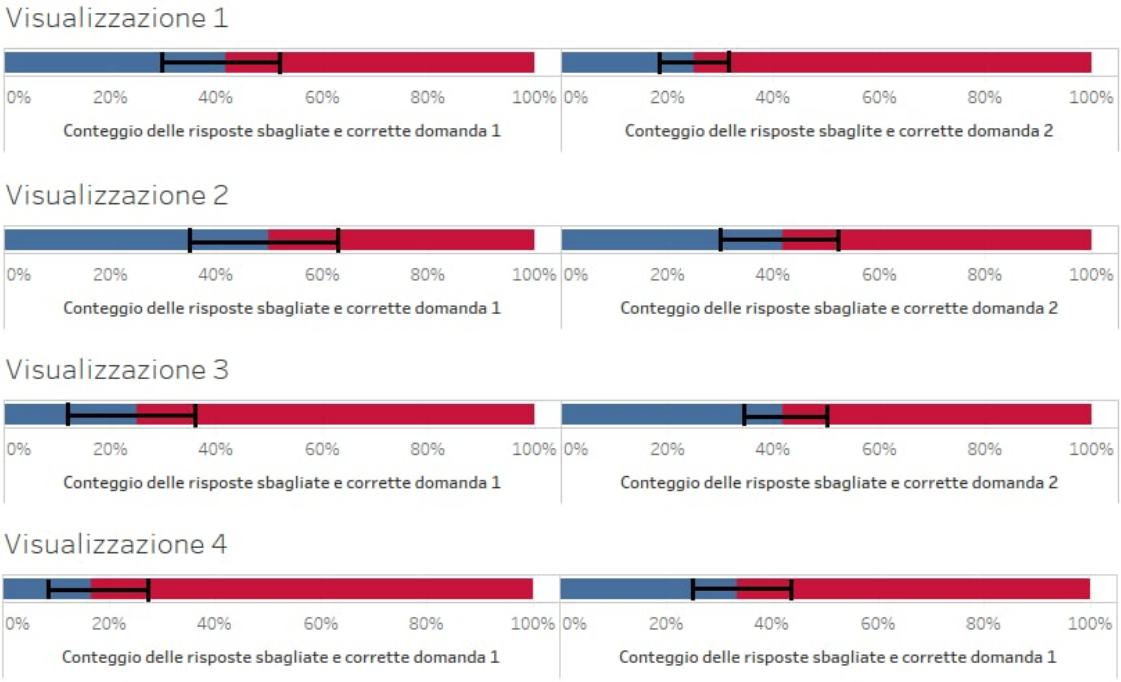


Figure 8: Percentuale di risposta alle domande

4.2 Test euristico

Sono state intervistate 6 persone per osservare il loro approccio nel visualizzare le infografiche costruite. Tutti gli intervistati, non esperti di settore, hanno avuto bisogno di tempo per relazionarsi con il software Tableau. In particolare, per la seconda e terza visualizzazione in un primo momento, non hanno notato sia i filtri che la doppia scala. Nella terza infografica, invece, L'andamento della sentiment analysis non è stata immediatamente compresa, anche perchè per la maggior parte degli intervistati è stata la prima volta ad interfacciarsi con questo tipo di analisi. L'ultima dashboard, invece, è stata efficace per quanto riguarda la correlazione tra le due variabili considerate. Si è deciso quindi di apportare leggere modifiche, in particolare aggiungendo commenti per poter inoltrare il lettore alla lettura corretta della visualizzazione.

4.3 Questionario Psicometrico

Sono state intervistate 24 persone per il questionario psicometrico utilizzando la scala Cabitza - Locoro. Si presentano graficamente i risultati:

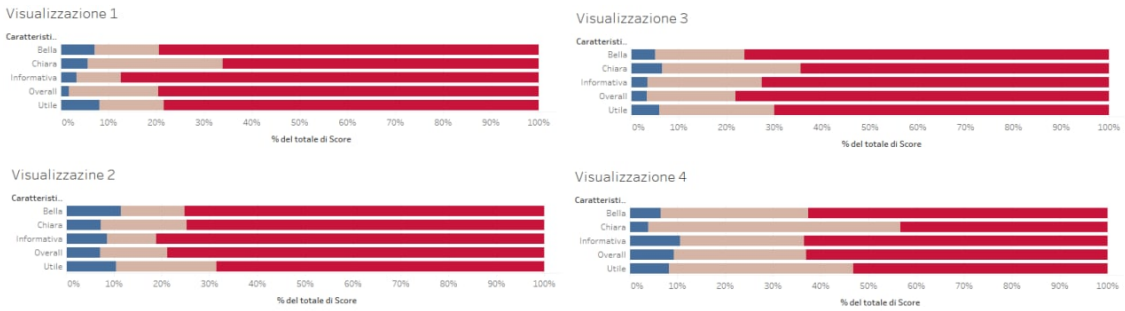


Figure 9: Percentuale di Risposta al Questionario Psicometrico

Le valutazioni sono quasi tutte positive, anche se la maggior parte delle risposte si assestano tra un punteggio di 4 e 5. Le valutazioni degli intervistati con conoscenze statistiche sono state quasi sempre superiori per la facilità di comprensione dei grafici proposti. Di seguito, si riporta il corrplot delle metriche utilizzate:

Figure 10: Matrice di Correlazione delle metriche del questionario

5 Conclusioni

Per concludere, le visualizzazioni hanno avuto lo scopo di mostrare se l'andamento del mercato azionario viene influenzato dal pensiero del pubblico. Nonostante il fatto che alcune notizie sembrano corrispondere a particolare picchi nel numero di tweet e nei prezzi azionario, verificare se vi sia una correlazione tra i due andamenti richiederebbe di integrare il lavoro con uno studio analitico e statistico. Dagli scatter plot si è visto come la correlazione tra il numero di tweet e il prezzo azionario sembra dipendere dall'azienda considerata. Per uno sviluppo futuro, si potrebbe implementare un algoritmo specifico per quanto riguarda un'analisi in grado di determinare il pensiero collettivo riguardante le aziende considerate, il quale sarebbe dunque in grado di determinare in quale condizione si trova questa "mente corale" e quindi effettuare una previsione rialzista o ribassista sul valore dei mercati finanziari.